DUALTRACE: PROBING EVIDENCE INTEGRATION FOR RAG HALLUCINATION DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Retrieval-augmented language models can rely on retrieved text yet still misrepresent it: context dependence alone does not establish faithfulness. We introduce DualTrace, a supervised detector built around two complementary views of evidence integration: how removing context changes a fixed response computation, and whether predictions decoded from hidden states agree across depth. DualTrace replays the same response with and without context, measuring output changes, hidden-state movement, and intermediate-to-final prediction disagreement, then pools these signals for response-level classification. On RAGTruth and HalluRAG, DualTrace improves on the evaluated final-layer controls and transfers across datasets and models. Averaging feature-family detector scores also outperforms selecting a single family, supporting their complementary predictive value. Across settings, hallucinated responses exhibit lower early uncertainty but greater late-layer disagreement. Context sensitivity further correlates with fixed-response log-probability changes under activation replacement. Together, these results connect a practical detection representation with measurable changes in evidence-sensitive computation and prediction agreement across depth.

1 INTRODUCTION

Retrieval-augmented generation (RAG) supplies a language model with evidence beyond its stored knowledge (Lewis et al., 2020), yet the response may still introduce unsupported details or contradict the documents (Niu et al., 2024; Ridder & Schilling, 2025). A faithful answer may change little without retrieval when the model already knows the answer; conversely, an unfaithful answer may depend strongly on retrieved text while misinterpreting it. Context dependence therefore does not, by itself, establish faithfulness. This gap motivates examining both how evidence changes computation and whether predictions remain consistent across depth.

Prior internal-state detectors characterize different aspects of evidence use. ReDeEP measures context retention and feedforward contributions (Sun et al., 2025); LUMINA combines relevant-versus-random-context distributional differences with prediction changes across layers (Yeh et al., 2026); and CORTEX uses reference-induced activation differences for token-level prediction (Furumai et al., 2026). These approaches motivate looking beyond the final vocabulary distribution, which cannot reveal how context changes hidden computation. Comparing matched computations with and without context provides this layer-by-layer view.

We organize detection around two questions: how does removing context change the computation of a fixed response, and do predictions agree across layers when context is present? The first compares output behavior and hidden-state movement across paired computations. The second decodes intermediate states into token predictions and compares them with a final-layer reference. Because a prediction can be confident yet disagree with that reference, uncertainty and cross-layer agreement are not interchangeable. DualTrace combines these two diagnostic axes—context sensitivity and prediction agreement—for response-level faithfulness detection.

DualTrace is a supervised white-box detector that uses labeled responses while keeping the language model frozen. It evaluates the same observed response with retrieved context (CTX) and without it (NOCTX), so the two paths differ in evidence rather than generated content. Three feature families instantiate the two diagnostic axes: *Conflict* measures output changes and document alignment; *Drift* measures how hidden-state movement changes between paths; and *Instability* measures prediction

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

disagreement across CTX layers. We pool token-level measurements over the response and train XGBoost for classification.

Our main method contribution is this paired representation of evidence-sensitive computation and cross-layer prediction agreement. On RAGTruth and HalluRAG, it improves on the evaluated final-layer controls and transfers across datasets and models. Averaging separately trained feature-family scores also improves on the training-selected single family in aggregate, indicating that the families carry complementary predictive information.

We then analyze what these measurements reveal about evidence integration. In aggregate, hallucinated responses show lower early uncertainty but greater late-layer disagreement. Hallucination rates also vary across combinations of context sensitivity and prediction disagreement, with patterns that depend on the model and dataset. Finally, replacing NOCTX hidden states with CTX states links context sensitivity to changes in fixed-response log probability. These analyses connect the detector's signals to both hallucination labels and measurable responses to activation replacement.

2 Related Work

Hallucination detectors differ mainly in what evidence they observe. SelfCheckGPT measures consistency across sampled responses (Manakul et al., 2023), while reference-based methods such as AlignScore compare generated text with supporting content (Zha et al., 2023). Internal probes instead use model activations or attention: Factoscope studies activation and token-prediction patterns (He et al., 2024), Lookback Lens uses context-to-generation attention ratios (Chuang et al., 2024a), and semantic entropy probes predict sampling-derived uncertainty from a single generation (Kossen et al., 2024). RAGLens learns an additive detector over selected sparse-autoencoder features (Xiong et al., 2026). DualTrace differs by replaying the same response with and without context and classifying the resulting paired computation.

The closest methods differ in how they measure context effects and intermediate computation. ReDeEP constructs external-context and parametric-knowledge scores from attention heads and feedforward updates (Sun et al., 2025). LUMINA combines a relevant-versus-random-context distributional distance with a cross-layer information-processing score (Yeh et al., 2026). CORTEX combines context-induced activation differences with preceding-token information and smooths token-level predictions (Furumai et al., 2026). DualTrace instead pairs context-removal effects on outputs and hidden-state movement with intermediate-to-final prediction disagreement. We compare this representation both with each method's native scoring rule or classifier and with XGBoost adaptations of its features.

Cross-layer comparisons have also been used for generation in DoLa (Chuang et al., 2024b), while context-aware decoding contrasts predictions with and without context (Shi et al., 2024). HARP uses output-head subspace projections for detection (Hu et al., 2026); DualTrace includes related summaries within Drift. Computational-graph verification links diagnostic representations to targeted activation changes (Zhao et al., 2026). Following this motivation, we test whether DualTrace's measurements track fixed-response log-probability changes under hidden-state replacement.

3 Detection with Paired Internal States

3.1 Response-aligned computation

Given a query q, retrieved documents d, and an observed response $r = (y_1, \ldots, y_T)$, DualTrace predicts a response-level hallucination label $z \in \{0, 1\}$ (Figure 1). A positive label indicates that the response contains hallucinated content under the benchmark's evidence and annotation rules. The language model remains frozen; only the detector is trained from response-level labels.

Generating separate responses would confound context changes with changes in answer content. We therefore replay the same observed tokens on both paths. Under teacher-forced replay, the fixed response is evaluated with $x^{\mathrm{CTX}} = (q, d)$ and $x^{\mathrm{NOCTX}} = q$, producing hidden states $h^p_{t,\ell}$ and vocabulary logits a^p_t for path $p \in \{\mathrm{CTX}, \mathrm{NOCTX}\}$. Here t indexes the position after consuming y_t, and ℓ indexes the returned hidden-state sequence. Thus a^p_t predicts the next token given $(x^p, y_{\leq t})$.

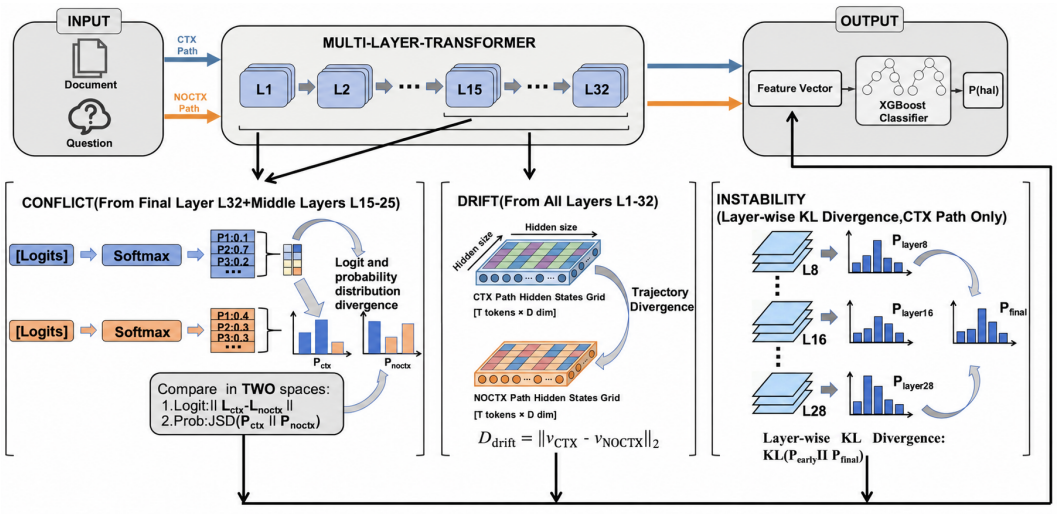

Figure 1: DualTrace measures context effects with Conflict and Drift, and cross-layer prediction disagreement with Instability. Both paths evaluate the same response. Pooled features feed a supervised XGBoost classifier. Layer numbers illustrate a 32-layer model; v contains the movement magnitudes in Equation 3.

Target-token likelihoods in the output and replay controls are read instead from the preceding causal position.

For a response token, its internal trajectory is the sequence of hidden states across layers. Because the two paths share the model and observed response, context removal is the only input change. Each path requires one forward pass; when contextual generation states are cached, detection adds only the NOCTX pass.

3.2 Three feature families for the two questions

Conflict and Drift capture context sensitivity, while Instability captures cross-layer prediction disagreement. We define representative measurements below; Appendix A lists the full feature inventory.

Conflict asks how strongly the output changes when retrieval is removed, while retaining document-alignment cues that help interpret that change. Let $P_t^p = \mathrm{softmax}(a_t^p)$, with $P_t^p(v) = \exp(a_t^p(v))/\sum_{u\in\mathcal{V}}\exp(a_t^p(u))$. Two representative summaries are the vocabulary-logit distance and Jensen–Shannon divergence (JSD):

$$C_{\mathrm{logit}}(t) = \|a_t^{\mathrm{CTX}} - a_t^{\mathrm{NOCTX}}\|_2, \tag{1}$$

$$C_{\mathrm{JSD}}(t) = \tfrac{1}{2}\,\mathrm{KL}(P_t^{\mathrm{CTX}}\|M_t) + \tfrac{1}{2}\,\mathrm{KL}(P_t^{\mathrm{NOCTX}}\|M_t), \quad M_t = \tfrac{1}{2}(P_t^{\mathrm{CTX}} + P_t^{\mathrm{NOCTX}}). \tag{2}$$

Here $\mathrm{KL}(P\|Q) = \sum_{v\in\mathcal{V}} P(v)\log[P(v)/Q(v)]$ uses natural logarithms. The logit distance preserves scale differences, whereas JSD measures a symmetric change bounded by $\log 2$. Neither alone implies contradiction; both measure context sensitivity. Cosine alignment with document states, final-state differences, and path-specific norms complete the family.

Output changes can miss substantial internal differences. Drift therefore asks whether retrieval changes how the hidden state moves through depth. Define the displacement $\Delta h_{t,\ell}^p = h_{t,\ell+1}^p - h_{t,\ell}^p$, its magnitude $v_{t,\ell}^p = \|\Delta h_{t,\ell}^p\|_2$, and the vector $v_t^p = (v_{t,0}^p, \ldots, v_{t,L-1}^p)$. The trajectory-divergence feature compares these layerwise movement magnitudes:

$$D(t) = \|v_t^{\mathrm{CTX}} - v_t^{\mathrm{NOCTX}}\|_2 = \sqrt{\sum_{\ell=0}^{L-1}\left(v_{t,\ell}^{\mathrm{CTX}} - v_{t,\ell}^{\mathrm{NOCTX}}\right)^2}. \tag{3}$$

The distance grows when context changes the amount of hidden-state movement at corresponding layers. Within-path summaries retain the maximum, late-layer mean, total, and variance of movement

magnitudes, together with changes in direction. Additional features project states onto subspaces derived from the output head, following Hu et al. (2026); Appendix A defines these projections.

Instability addresses the second axis: whether predictions agree across depth when context is present. A *logit lens* decodes selected hidden states into vocabulary distributions using the model's final normalization $\mathcal{N}$ and output head W_U:

$$\widetilde{P}_{t,\ell} = \mathrm{softmax}\big(W_U \mathcal{N}(h_{t,\ell}^{\mathrm{CTX}})\big)\,, \qquad I_\ell(t) = \mathrm{KL}\Big(\widetilde{P}_{t,\ell} \| \widetilde{P}_{t,L}\Big)\,. \tag{4}$$

Applying the lens to the final returned state defines the *final-layer lens prediction*, which is distinct from the model's direct output distribution (Appendix A). We retain early, middle, and late KL, early, middle, and final entropy, and top-prediction agreement. Entropy measures confidence at a layer; KL measures disagreement with the final-layer reference. Retaining both separates local certainty from consistency across depth.

3.3 Response pooling and classification

A response-level detector must preserve both global behavior and localized peaks. For each token-feature sequence $X_{:,j}$, we therefore retain its mean, standard deviation, maximum, and mean of the K largest values (Top-K mean). The tail size is $K = \min(T, \max(5, \lceil 0.1T \rceil))$. With $\mathcal{T}_j$ denoting the indices of the K largest values in dimension j, the pooled representation is

$$\phi_j(r) = \left[\frac{1}{T}\sum_t X_{t,j},\ \sqrt{\frac{1}{T}\sum_t (X_{t,j} - \overline{X}_j)^2},\ \max_t X_{t,j},\ \frac{1}{K}\sum_{t \in \mathcal{T}_j} X_{t,j} \right]. \tag{5}$$

The mean and standard deviation summarize the full response, while the maximum and Top-K mean retain localized peaks that may occur at different positions for different features (Appendix A.1). Four statistics for 36 token features plus response length yield a 145-dimensional vector.

XGBoost (Chen & Guestrin, 2016) combines differently scaled features and learns their joint relation to hallucination. Cross-validation groups responses that share a retrieval source, selects boosting rounds, and produces out-of-fold (OOF) scores from models trained without each response's fold. We then refit on the full training split for five seeds and average test scores. The F1-maximizing threshold is chosen from mean training-OOF scores; test labels are used only for evaluation (Appendix B).

To test complementarity without learning combination weights, we also average scores from separately trained feature-family detectors. If $s_C(r), s_D(r), s_I(r)$ are their scores, then

$$s_{\mathrm{fusion}}(r) = \tfrac{1}{3}\big(s_C(r) + s_D(r) + s_I(r)\big). \tag{6}$$

This equal-weight fusion combines three supervised family detectors without learning an additional combiner. Comparing it with the family selected by training-OOF AUROC tests whether fixed score averaging adds predictive value beyond choosing one family.

4 Detection Performance and Feature Comparisons

4.1 Datasets and evaluation protocols

We evaluate on RAGTruth's annotated RAG responses (Niu et al., 2024) and HalluRAG's closed-domain responses (Ridder & Schilling, 2025). The broad benchmark covers Llama2-7B-Chat, Llama2-13B-Chat, Llama3-8B-Instruct, and Mistral-7B-Instruct-v0.1; the targeted output-control study includes Qwen3-8B and Qwen3-14B. We use three evaluation sets: the benchmark splits (Appendix C.1), control subsets with all required features available, and matched response sets for competing methods. The 12 control subsets contain 488–798 training and 106–188 test responses, with disjoint retrieval-source groups (Appendix B). Each method comparison uses responses available to every method in that comparison; the representation analysis uses its specified shared set.

We report AUROC and AUPRC, treating hallucination as the positive class. The broad comparison also reports Pearson correlation (PCC) with the fraction of hallucinated response tokens. In the targeted evaluation, all DualTrace variants share five seeds, grouped training-OOF selection, pooling, and XGBoost settings; scores within a setting average the five predictions. Macro averages weight

Table 1: Broad response-level comparison on the benchmark splits. AUC: AUROC; AP: AUPRC; PCC: correlation with the fraction of hallucinated tokens (all higher-is-better). Bold marks the numerical best within each model–dataset–metric column. The output-control and matched-method comparisons use separately specified response sets (Tables 2 and 3).

Method	Llama2-7B			Llama2-13B			Llama3-8B			Mistral-7B		
	AUC	AP	PCC	AUC	AP	PCC	AUC	AP	PCC	AUC	AP	PCC
RAGTruth												
Perplexity	0.5307	0.6107	-0.0644	0.4230	0.4885	-0.1823	0.8189	0.7003	0.0350	0.5995	0.6891	0.0517
LN-Entropy	0.6702	0.7370	**0.2146**	0.7146	0.7160	**0.3525**	0.8076	0.6843	**0.4581**	0.7462	0.8024	**0.3158**
Focus	0.6144	0.6782	0.1025	0.5902	0.5027	0.1032	0.7326	0.4625	0.2890	0.6988	0.6557	0.2420
SelfCheckGPT	0.5970	0.5734	0.0327	0.6248	0.6123	-0.0076	0.6039	0.6010	-0.0431	0.6474	0.6288	0.0713
EigenScore	0.5867	0.5527	-0.0685	0.7929	0.7793	0.1692	0.8049	0.7286	0.2676	0.7378	0.7510	0.0594
P(True)	0.6509	0.7398	0.1990	0.5413	0.4900	0.0410	0.7415	0.5832	0.3690	0.4241	0.4238	-0.0387
RefChecker	0.5951	0.6189	0.0260	0.7774	0.7770	0.1497	0.8146	0.7909	0.3295	0.6576	0.6748	0.1512
AlignScore	0.6342	0.6015	0.0954	0.6517	0.6234	0.1087	0.6894	0.6582	0.1342	0.6428	0.6120	0.1021
ReDeEP	0.7522	0.7528	0.1036	0.8311	0.8363	0.1959	0.8677	0.8496	0.3685	0.8271	0.8357	0.1898
DualTrace	**0.7947**	**0.8133**	0.2043	**0.8578**	**0.8558**	0.2281	**0.9354**	**0.8849**	0.3849	**0.8448**	**0.8641**	0.2162
HalluRAG												
Perplexity	0.5124	0.5210	-0.0123	0.8462	0.8719	0.2553	0.8214	0.8342	0.1523	0.5841	0.5923	0.0421
LN-Entropy	0.6432	0.6120	0.1842	0.6397	0.7135	0.2627	0.7915	0.7841	0.3842	0.7215	0.7341	0.2910
Focus	0.6486	0.6973	0.2541	0.6592	0.7228	0.2662	0.5442	0.8032	-0.0379	0.6676	0.7245	0.2874
SelfCheckGPT	0.6066	0.4787	0.0147	0.6392	0.6148	0.2515	0.5964	0.5979	0.1923	0.7332	0.6285	0.4896
EigenScore	0.8238	0.8031	0.5276	0.7148	0.7825	0.3424	0.6708	0.3845	0.2777	0.6200	0.6532	0.1984
P(True)	0.7484	0.7212	0.3625	0.7912	0.7877	**0.5469**	0.7399	0.8070	0.2094	0.8720	0.8833	0.5666
RefChecker	0.5789	0.4455	-0.0876	0.7659	0.7802	0.3221	0.6899	0.7081	0.4212	0.7448	0.7512	0.1255
AlignScore	0.7188	0.6723	0.2145	0.7412	0.7029	0.2312	0.7725	0.7381	0.2687	0.7256	0.6854	0.2238
ReDeEP	0.9248	0.9266	0.5059	0.8326	0.8452	0.4481	0.8331	0.8111	0.4411	0.8871	0.8954	0.5892
DualTrace	**0.9582**	**0.9573**	**0.6880**	**0.8655**	**0.9027**	0.5067	**0.8489**	**0.8410**	**0.5840**	**0.9246**	**0.9416**	**0.6681**

settings equally. Confidence intervals (CIs) use 10,000 bootstrap repetitions over retrieval-source groups while preserving paired comparisons. Holm correction covers the comparison family for each analysis, and detector thresholds are calibrated only from training predictions.

4.2 Broad benchmark comparison

We first compare detectors spanning uncertainty, sampled answers, self-assessed truth, textual support, and internal states. Table 1 reports nine baselines; Appendix C.1 details their evaluation.

DualTrace has the highest AUROC and AUPRC in all eight broad-benchmark settings. AUROC ranges from 0.7947 to 0.9354 on RAGTruth and from 0.8489 to 0.9582 on HalluRAG. On Llama3-8B/RAGTruth, for example, it reaches 0.9354 versus 0.8677 for ReDeEP and 0.8189 for Perplexity. We next isolate the value of internal paired traces using final-layer controls and matched comparisons with the closest internal-state methods.

4.3 Output controls and matched-method comparisons

We compare the full detector with two final-layer controls that keep the training pipeline fixed. Output-gap uses nine final-layer differences, including target logit, log probability, entropy, rank, JSD, and symmetric KL. Output-all adds ten path-specific values such as target likelihood and entropy. Both use the same responses, splits, pooling, length feature, and XGBoost classifier as DualTrace. Output controls use positions that predict the observed token, whereas DualTrace uses states after consuming it. Appendix A.2 lists the complete controls and position conventions.

Full DualTrace exceeds both output controls in all 12 point estimates (Table 2). Macro AUROC is 0.9279, compared with 0.8825 for Output-all and 0.8766 for Output-gap. The mean gain over Output-all is 0.0454; macro AUPRC is 0.8995 versus 0.8300. These results show that paired internal traces provide response-level ranking information beyond the evaluated likelihood and final-distribution summaries.

Table 2: Response-level AUROC on the 12 control-subset settings. Output-all uses final-layer values and differences; Output-gap uses differences only. The last column is the 95% source-group bootstrap interval for Full minus Output-all. Scores are five-seed prediction ensembles; the macro weights settings equally.

Model	Dataset	Full	Output-all	Output-gap	Δ 95% CI
Llama2-7B	HalluRAG	0.9827	0.9360	0.9334	$[-0.001, 0.105]$
Llama2-7B	RAGTruth	0.8789	0.7523	0.7615	$[0.040, 0.216]$
Llama2-13B	HalluRAG	0.8865	0.8473	0.8412	$[-0.035, 0.117]$
Llama2-13B	RAGTruth	0.8970	0.8736	0.8559	$[-0.040, 0.085]$
Llama3-8B	HalluRAG	0.9472	0.8851	0.8769	$[-0.009, 0.138]$
Llama3-8B	RAGTruth	0.9884	0.9355	0.9246	$[0.016, 0.099]$
Mistral-7B	HalluRAG	0.9627	0.9599	0.9577	$[-0.038, 0.048]$
Mistral-7B	RAGTruth	0.8676	0.8208	0.8343	$[-0.010, 0.104]$
Qwen3-8B	HalluRAG	0.9029	0.8679	0.8736	$[-0.054, 0.126]$
Qwen3-8B	RAGTruth	0.9382	0.8775	0.8519	$[0.017, 0.108]$
Qwen3-14B	HalluRAG	0.9450	0.9048	0.8803	$[-0.043, 0.126]$
Qwen3-14B	RAGTruth	0.9384	0.9297	0.9283	$[-0.026, 0.047]$
Macro		0.9279	0.8825	0.8766	—

Table 3: Matched-method AUROC on the unified RAGTruth evaluation sets. Full DualTrace: L3 (Llama3-8B) 0.8924; L2 (Llama2-7B) 0.8085. Intervals are Full minus competitor; p_{Holm} adjusts all seven AUROC comparisons. Sup.: detector supervision. [a]Official features, paper scoring, train-only calibration. [b]Paper-equivalent memory-bounded implementation. [c]Paper reimplementation with its MLP. [d]Official SAE features and GAM. XGB rows share DualTrace's classifier protocol. Appendix C.3 gives details.

Model	Method / classifier	Sup.	AUROC	Δ 95% CI	p_{Holm}
L3	ReDeEP native[a]	Calibration	0.7850	$[0.0707, 0.1449]$	0.0014
L3	ReDeEP + XGB	Response	0.8740	$[-0.0037, 0.0413]$	0.5100
L3	LUMINA native[b]	None	0.8751	$[-0.0070, 0.0424]$	0.6712
L3	LUMINA + XGB	Response	0.8930	$[-0.0188, 0.0181]$	1.0000
L3	CORTEX[c]	Token	0.9185	$[-0.0498, -0.0025]$	0.1830
L2	RAGLens native[d]	Response	0.8250	$[-0.0465, 0.0135]$	0.8430
L2	RAGLens + XGB	Response	0.8010	$[-0.0235, 0.0385]$	1.0000

We next compare alternative uses of internal states on matched RAGTruth responses (Table 3). ReDeEP uses its official features with the paper scoring rule and train-only calibration; LUMINA uses its original scoring rule with a paper-equivalent memory-bounded implementation; CORTEX uses a paper reimplementation; and RAGLens uses official sparse-autoencoder (SAE) features with a generalized additive model (GAM). We also evaluate ReDeEP, LUMINA, and RAGLens features under DualTrace's XGBoost protocol. Each method's original rule or classifier is labeled *native*; the adaptation is labeled *+XGB*. DualTrace uses response labels, whereas CORTEX uses token labels. Appendix C.3 gives implementation and supervision details.

DualTrace reaches 0.8924 AUROC on Llama3-8B, exceeding native ReDeEP by 0.1074 ($p_{\mathrm{Holm}} = 0.0014$). None of the other six differences is significant after the seven-comparison Holm correction. LUMINA+XGB is numerically similar at 0.8930, while CORTEX and native RAGLens have higher point estimates on their respective backbones. We therefore interpret these comparisons as competitive rather than uniformly superior performance.

4.4 Signal complementarity and evidence-path controls

Family removal tests how much each signal family contributes to the joint detector. Equal-weight score averaging separately tests whether combining family predictions improves over selecting one family from training data.

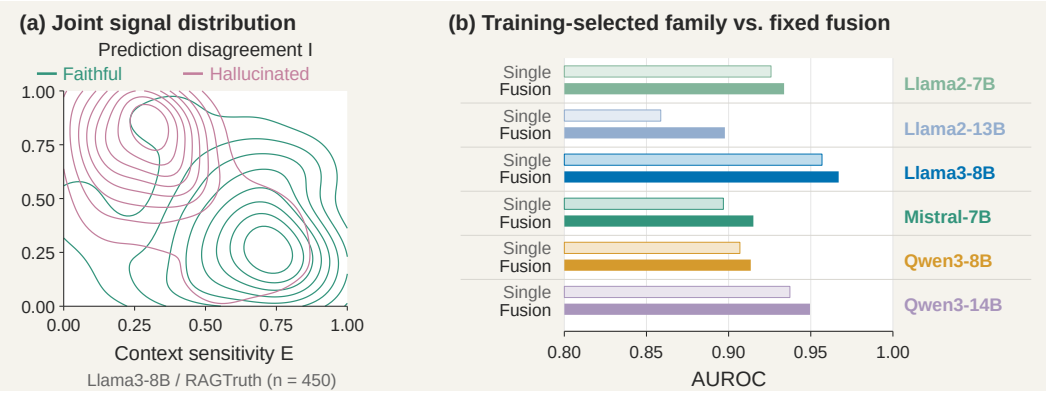

Figure 2: Signals and their combination. (a) Class-conditional density contours for 450 Llama3-8B/RAGTruth responses. Context sensitivity E combines output JSD and trajectory divergence; prediction disagreement I measures late-to-final KL (Equation 7). (b) AUROC averaged equally over the two datasets for each model: training-OOF-selected single family versus equal-weight fusion of supervised family scores. The AUROC axis starts at 0.80.

Table 4: Transfer AUROC with source-only training and calibration. Entries average per-seed AUROC over five seeds, then weight directions equally. Cross-model transfer retains the dataset and holds out target test sources.

Transfer axis	Full	Compact	Output-all
Cross-dataset (12)	0.9220	0.9077	0.8713
Cross-model (6)	0.9203	0.9081	0.8985
All directions (18)	0.9215	0.9078	0.8804

Conflict and Drift reach macro AUROCs of 0.9176 and 0.9111, while Instability reaches 0.8435. Full improves on each family in aggregate (Appendix C.2). Removing Conflict, Drift, or Instability yields 0.9161, 0.9200, and 0.9245, respectively. The relative contribution varies by family and setting, but the aggregate results favor retaining both context-sensitive and cross-layer signals.

Equal-weight score averaging reaches 0.9294 AUROC, compared with 0.9137 for the single family chosen from training-OOF scores (Figure 2b), and is higher in 11 of 12 settings. The 0.0157 macro gain supports complementary information across family scores. Full joint training and equal-weight fusion each have the higher point estimate in six settings. Appendix C.2 provides the complete results and finer-grained ablations.

To test whether these gains depend on the retrieved evidence, we vary the supplied documents while keeping responses and labels fixed. Across six settings with responses shared across conditions, Full-minus-Output-all gains average 0.0273 for CTX/NOCTX, 0.0381 for CTX/random-context, and 0.0322 for CTX+random-context/NOCTX. Full is higher in 16 of the 18 comparisons that include the original evidence, versus two of six random/NOCTX comparisons. The gains are therefore more consistent when one path contains the retrieved evidence. Appendix C.4 reports the full four-condition matrix. Appendix C.5 separately tests replay-likelihood artifacts and faithful responses whose natural NOCTX answers agree with the supported answers.

4.5 Transfer and Computational Cost

To test reuse without target labels, we train and calibrate on one model–dataset setting (source) and evaluate on another (target). We consider both dataset directions for each model and six same-dataset cross-model directions, keeping source groups disjoint. Full DualTrace reaches mean AUROC 0.9215 across 18 directions, compared with 0.8804 for Output-all (Table 4); target labels are used only for evaluation. The Compact subset, with 11 rather than 36 features, reaches 0.9078. These results support reuse of the learned representation and classifier across the evaluated settings (Appendix C.6).

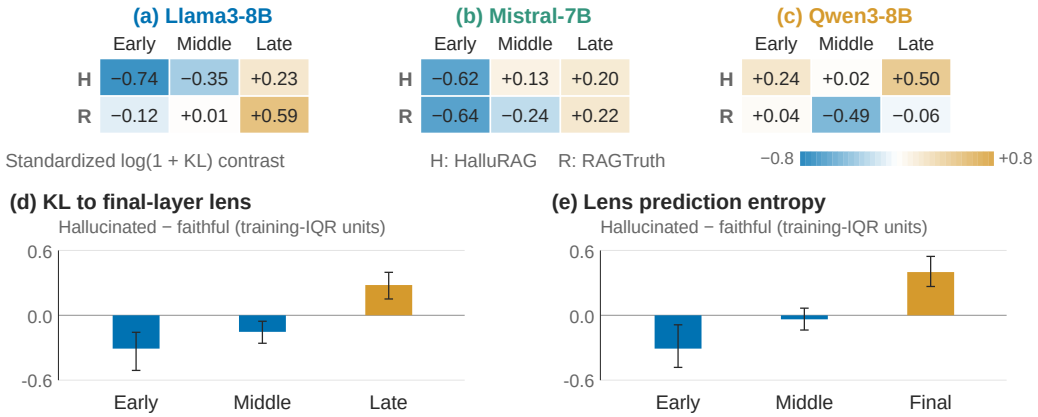

Figure 3: Prediction differences across depth, hallucinated minus faithful. (a–c) KL contrasts for each model and dataset (H: HalluRAG; R: RAGTruth). (d–e) Six-setting equal-weight means with 95% source-group bootstrap intervals (10,000 repetitions). Values use training-IQR units, with $\log(1 + \mathrm{KL})$ applied before scaling. Early/middle/late use returned-state indices 8/20/28; final denotes the final-layer lens reference. Table 16 gives all measurements and corrected tests.

The extra NOCTX pass trades computation for an explicit context comparison. We measure forward-pass overhead separately from end-to-end detection cost. On 100 held-out responses per model, median dual-path/CTX-output-only forward-time ratios are 1.388, 1.327, and 1.362 for Llama3-8B, Mistral-7B, and Qwen3-8B. A separate full-detector benchmark includes tokenization, both forwards, feature computation, pooling, and a 100-round XGBoost classifier; median times are 0.230, 0.233, and 0.281 seconds per response. Appendix C.7 defines the timing boundaries.

5 Context Sensitivity and Prediction Agreement

5.1 Joint Analysis of Selected Measurements

We next examine three representative measurements underlying the detector: output JSD and trajectory divergence for context sensitivity, and late-to-final lens KL for prediction disagreement. On the six-setting CTX/NOCTX evidence-path data, we average each quantity over response positions and map it to a training-relative percentile. With percentile functions $\widehat{F}$, context sensitivity E and prediction disagreement I are

$$E(r) = \tfrac{1}{2}\left[\widehat{F}_C(\overline{C}_{\mathrm{JSD}}(r)) + \widehat{F}_D(\overline{D}(r))\right], \qquad I(r) = \widehat{F}_I(\overline{I}_{\mathrm{late}}(r)). \tag{7}$$

Figure 2a illustrates the joint distribution on Llama3-8B/RAGTruth. Fixed cutoffs at 0.5 define four descriptive groups for comparing hallucination rates. The full analysis contains 1,671 held-out responses from Llama3-8B, Mistral-7B, and Qwen3-8B across both datasets.

On RAGTruth, the low-E/high-I group has a higher hallucination rate than the high-E/low-I group by 64.06, 21.56, and 25.74 percentage points for Llama3-8B, Mistral-7B, and Qwen3-8B, respectively (Figure 8). The six-setting macro difference is 17.75 points (95% CI [10.35, 25.20], Holm-adjusted $p = 0.0008$). HalluRAG shows a different pattern for Llama3-8B, with higher rates in the high-E groups. Context sensitivity therefore cannot be interpreted as a universal monotonic risk score; it must be read jointly with prediction disagreement and the model–dataset setting. This variation motivates a learned classifier rather than a fixed high-sensitivity rule. Appendix E gives all group contrasts.

5.2 Prediction Consistency Across Depth

To examine how prediction behavior changes with depth, we compare early, middle, and late CTX lens predictions. Each response-level measurement is scaled using training medians and interquartile ranges (IQRs), after the transformations in Appendix E.

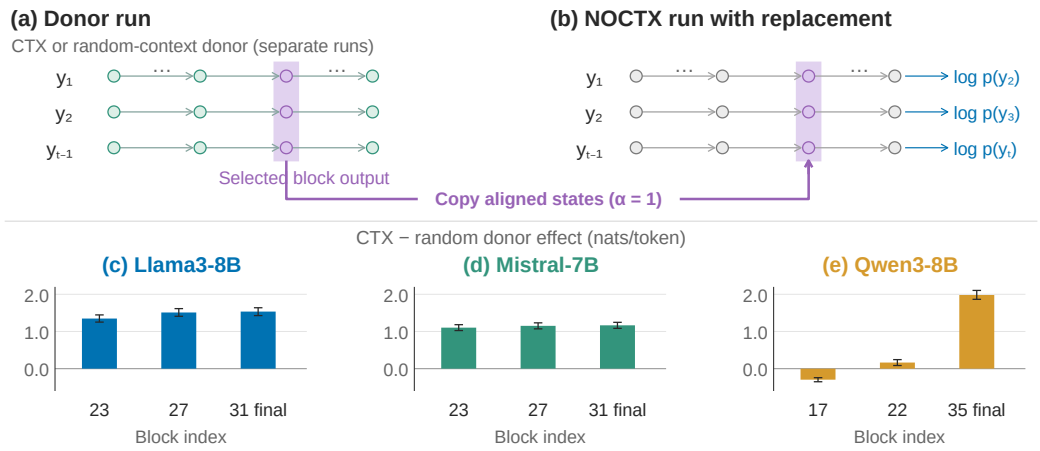

Figure 4: Activation replacement and its effect on fixed-response probability. (a–b) Donor block outputs replace aligned NOCTX states at positions $1, \ldots, T-1$ ($\alpha = 1$), scoring tokens $2, \ldots, T$. CTX and random-context donors run separately. (c–e) CTX-minus-random effects in nats/token with 95% source-group bootstrap intervals (10,000 repetitions). Blocks are zero-based and discovery-selected; final-block replacement is a readout reference. The three models contribute 120, 119, and 120 responses.

Averaged equally across six settings, hallucinated responses show lower early entropy and smaller early-to-final KL, but larger late-to-final KL and higher final entropy (Figure 3). The hallucinated-minus-faithful contrast in late-minus-early log-KL, defined as $\log(1 + I_{\text{late}}) - \log(1 + I_{\text{early}})$ before standardization, is positive in five settings. An early probe can therefore appear confident and relatively aligned while later predictions diverge, which motivates retaining both uncertainty and cross-layer agreement signals.

5.3 Response probability under activation replacement

We finally ask whether context sensitivity tracks a measurable change in fixed-response probability. On 359 held-out RAGTruth responses, we replace NOCTX states with $h' = (1-\alpha)h^{\text{NOCTX}} + \alpha h^{\text{CTX}}$, using layers selected on separate discovery data. CTX supplies the replacement states (the *donor*), while random-context states provide controls. At $\alpha = 1$, $J(r)$ averages the CTX-minus-random effect on mean response log probability across selected layers (Appendix D).

Context sensitivity E correlates with this log-probability effect at $\rho = 0.625$ (three-model macro; 95% CI [0.549, 0.691], Holm-adjusted $p = 0.0006$). Prediction disagreement I is negatively associated with the effect ($\rho = -0.308$, 95% CI $[-0.399, -0.209]$, $p_{\text{Holm}} = 0.0006$). Thus, higher E accompanies a larger signed response to CTX replacement, whereas higher I accompanies a smaller one.

Both the replacement effects and their associations with context sensitivity vary across blocks (Figure 4c–e). Averaged over the two selected non-final blocks, the correlation with context sensitivity is 0.797 for Llama3-8B and 0.890 for Mistral-7B. For Qwen3-8B, mean CTX-minus-random effects are -0.296, 0.166, and 1.981 at blocks 17, 22, and 35, with correlations -0.423, -0.294, and 0.788. The probability response associated with context sensitivity is therefore model- and layer-dependent; Appendix E gives aggregate correlation intervals.

6 Conclusion

DualTrace detects RAG hallucinations by combining evidence-sensitive changes in computation with prediction agreement across depth. Benchmark, output-control, score-averaging, and transfer results support this paired representation. Across settings, hallucinated responses show lower early uncertainty but greater late-layer disagreement, while context sensitivity tracks fixed-response probability changes under activation replacement.

AI USE STATEMENT

Generative AI tools assisted with refining research questions and hypotheses, designing control analyses, implementing experiment and analysis code, interpreting results, reviewing related literature, and drafting, translating, and editing the manuscript. They also assisted with preparing tables, correcting a formula in the framework schematic, and organizing the experimental artifacts. The authors are responsible for the data, implementation, interpretations, and final content of this work.

REPRODUCIBILITY STATEMENT

Section 3 defines the representation and classifier. Appendix A specifies the complete feature inventory, Appendix B gives the training and evaluation protocol, and Appendix E specifies the joint-measurement, statistical, and activation-replacement analyses. The supplementary material provides the DualTrace implementation, the control-subset and transfer configuration, and instructions describing the experimental protocols and required external inputs.

REFERENCES

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. Inside: Llms' internal states retain the power of hallucination detection, 2024. URL https://arxiv.org/abs/2402.03744.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system, 2016. URL https://arxiv.org/abs/1603.02754.

Yung-Sung Chuang, Linlu Qiu, Cheng-Yu Hsieh, Ranjay Krishna, Yoon Kim, and James R. Glass. Lookback lens: Detecting and mitigating contextual hallucinations in large language models using only attention maps. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 1419–1436, Miami, Florida, USA, November 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.84. URL https://aclanthology.org/2024.emnlp-main.84/.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. Dola: Decoding by contrasting layers improves factuality in large language models, 2024b. URL https://arxiv.org/abs/2309.03883.

Kazuaki Furumai, Shuichiro Haruta, Kazunori Matsumoto, and Daisuke Kamisaka. CORTEX: Token-level hallucination detection in RAG via comparative internal representations, 2026. URL https://arxiv.org/abs/2606.31033.

Jinwen He, Yujia Gong, Zijin Lin, Cheng'an Wei, Yue Zhao, and Kai Chen. LLM factoscope: Uncovering LLMs' factual discernment through measuring inner states. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 10218–10230, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.608. URL https://aclanthology.org/2024.findings-acl.608/.

Junjie Hu, Gang Tu, ShengYu Cheng, Jinxin Li, Jinting Wang, Rui Chen, Zhilong Zhou, and Dongbo Shan. HARP: Hallucination detection via reasoning subspace projection. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2509.11536.

Xiangkun Hu, Dongyu Ru, Lin Qiu, Qipeng Guo, Tianhang Zhang, Yang Xu, Yun Luo, Pengfei Liu, Yue Zhang, and Zheng Zhang. Refchecker: Reference-based fine-grained hallucination checker and benchmark for large language models, 2024. URL https://arxiv.org/abs/2405.14486.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk,

Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language models (mostly) know what they know, 2022. URL https://arxiv.org/abs/2207.05221.

Jannik Kossen, Jiatong Han, Muhammed Razzak, Lisa Schut, Shreshth Malik, and Yarin Gal. Semantic entropy probes: Robust and cheap hallucination detection in llms, 2024. URL https://arxiv.org/abs/2406.15927.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf.

Andrey Malinin and Mark Gales. Uncertainty estimation in autoregressive structured prediction, 2021. URL https://arxiv.org/abs/2002.07650.

Potsawee Manakul, Adian Liusie, and Mark Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9004–9017, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.557. URL https://aclanthology.org/2023.emnlp-main.557/.

Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, KaShun Shum, Randy Zhong, Juntong Song, and Tong Zhang. RAGTruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10862–10878, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.585. URL https://aclanthology.org/2024.acl-long.585/.

Jie Ren, Jiaming Luo, Yao Zhao, Kundan Krishna, Mohammad Saleh, Balaji Lakshminarayanan, and Peter J. Liu. Out-of-distribution detection and selective generation for conditional language models, 2023. URL https://arxiv.org/abs/2209.15558.

Fabian Ridder and Malte Schilling. The hallurag dataset: Detecting closed-domain hallucinations in rag applications using an llm's internal states, 2025. URL https://arxiv.org/abs/2412.17056.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Wen-tau Yih. Trusting your evidence: Hallucinate less with context-aware decoding. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pp. 783–791, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-short.69. URL https://aclanthology.org/2024.naacl-short.69/.

Zhongxiang Sun, Xiaoxue Zang, Kai Zheng, Yang Song, Jun Xu, Xiao Zhang, Weijie Yu, Yang Song, and Han Li. Redeep: Detecting hallucination in retrieval-augmented generation via mechanistic interpretability, 2025. URL https://arxiv.org/abs/2410.11414.

Guangzhi Xiong, Zhenghao He, Bohan Liu, Sanchit Sinha, and Aidong Zhang. Toward faithful retrieval-augmented generation with sparse autoencoders. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=hgBZP67BkP.

Samuel Yeh, Sharon Li, and Tanwi Mallick. LUMINA: Detecting hallucinations in RAG system with context–knowledge signals. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2509.21875.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. AlignScore: Evaluating factual consistency with a unified alignment function. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11328–11348, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.634. URL https://aclanthology.org/2023.acl-long.634/.

Tianhang Zhang, Lin Qiu, Qipeng Guo, Cheng Deng, Yue Zhang, Zheng Zhang, Chenghu Zhou, Xinbing Wang, and Luoyi Fu. Enhancing uncertainty-based hallucination detection with stronger focus. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 915–932, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.58. URL https://aclanthology.org/2023.emnlp-main.58/.

Zheng Zhao, Yeskendir Koishekenov, Xianjun Yang, Naila Murray, and Nicola Cancedda. Verifying chain-of-thought reasoning via its computational graph. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/c7b5a35ea98b62512a869c19ea7b03cb-Paper-Conference.pdf.

A Feature Definitions and Implementation

The full feature inventory contains 13 Conflict features, 15 Drift features, and eight Instability features. Conflict and Drift characterize the context's effect on outputs and hidden states; Instability characterizes prediction disagreement across CTX layers. Section 3.2 defines representative measurements. The three-feature Core variant uses output JSD, trajectory divergence, and late-to-final KL, with the same four pooling operators and response-length feature.

Hidden-state index zero denotes the embedding output. The intermediate entries follow transformer blocks, and the final returned hidden state follows the model's final normalization. Displacement calculations in Equation 3 operate directly on these model-returned states. The lens applies the model's final normalization to each decoded state, including the already final-normalized final returned state used as the lens reference. For the evaluated language models using root-mean-square normalization (RMSNorm), this operation is $\mathcal{N}(h)_i = w_i h_i / \sqrt{d_h^{-1} \sum_j h_j^2 + \epsilon}$, where d_h is the hidden dimension, w_i is the learned weight, and ϵ is the model's numerical stabilizer. The lens reference and the model's direct output logits consequently have separate definitions.

Let H_d contain the final contextual states at retrieved-document positions. The document-alignment feature is the mean of the ten largest cosine similarities between $h_{t,L}^{\mathrm{CTX}}$ and rows of H_d, or all rows if fewer than ten exist. The path-agreement feature is $\cos(a_t^{\mathrm{CTX}}, a_t^{\mathrm{NOCTX}})$. Their difference and a width-five moving average of that difference are retained. Other Conflict features are $\|h_{t,L}^{\mathrm{CTX}} - h_{t,L}^{\mathrm{NOCTX}}\|_2$, the logit distance in Equation 1, and the two final-state norms.

For each hidden-state index $\ell \in \{15, \ldots, 25\}$, context alignment is the cosine between $h_{t,\ell}^{\mathrm{CTX}}$ and the mean of H_d. We retain its mean, maximum, and sample variance across these indices. The inventory contains the mean under two names, `icr_mean` and `context_alignment_mean`; both features enter the evaluated representation. These features measure hidden-state cosine alignment with document states. The remaining Conflict feature is output JSD.

For Drift, within-path displacement magnitudes are normalized by their response-position mean at each layer:

$$\widehat{v}_{t,\ell} = \frac{v_{t,\ell}^{\mathrm{CTX}}}{T^{-1} \sum_{s=1}^{T} v_{s,\ell}^{\mathrm{CTX}} + 10^{-8}}. \tag{8}$$

We retain the maximum, its zero-based displacement index, the mean over the last five displacement steps, the sum, and the sample variance across depth. Directional acceleration is $1 - \cos(\Delta h_{t,\ell}^{\mathrm{CTX}}, \Delta h_{t,\ell+1}^{\mathrm{CTX}})$, summarized by its maximum and last-five mean. The paired trajectory distance uses the unnormalized magnitudes in Equation 3. Its companion feature averages products of sample-standardized CTX and NOCTX magnitude profiles across depth, with 10^{-8} added to each standard deviation.

The remaining six Drift features summarize an output-head projection inspired by HARP (Hu et al., 2026). Let $W_U^\top W_U = V \operatorname{diag}(\lambda) V^\top$, ordered by decreasing nonnegative eigenvalue. The implemented cutoff uses the number of cumulative-energy fractions strictly below 0.95, clipped to $[1, d_h - 1]$, to define k. Write $P_S = V_{1:k} V_{1:k}^\top$ and $P_R = I - P_S$. We retain $\|P_R h^{\mathrm{CTX}}\|_2$, its ratio to $\|h^{\mathrm{CTX}}\|_2 + 10^{-8}$, $\|P_S h^{\mathrm{CTX}}\|_2$, the cosine between adjacent response positions in $P_R h^{\mathrm{CTX}}$, $\|P_R(h^{\mathrm{CTX}} - h^{\mathrm{NOCTX}})\|_2$, and $\cos(P_R h^{\mathrm{CTX}}, P_R h^{\mathrm{NOCTX}})$. The adjacent-position cosine is set to one at the first position. The Projection ablation evaluates this six-feature subset within DualTrace's pooling and supervised classifier.

The lens probe indices are $\{8, 12, 16, 20, 24, 28\}$ in the returned hidden-state sequence. Early, middle, and late KL use indices 8, 20, and 28, respectively. Early and middle entropy use indices 8 and 20; final entropy uses the final-layer lens reference. Prediction stability is the fraction of adjacent probe pairs with the same top-ranked vocabulary item. The first-top-prediction feature records the ordinal index of the first probe matching the final lens top prediction, or -1 if none does. Entropy and KL use natural logarithms, with probabilities clamped below at 10^{-10} in the lens calculation. The evidence-path extraction computes probability-space quantities in float32.

Full uses all 36 features. Compact retains `delta_t`, `residual_shift`, `logit_ctx_noctx_jsd`, `icr_mean`, `late_velocity`, `trajectory_divergence`,

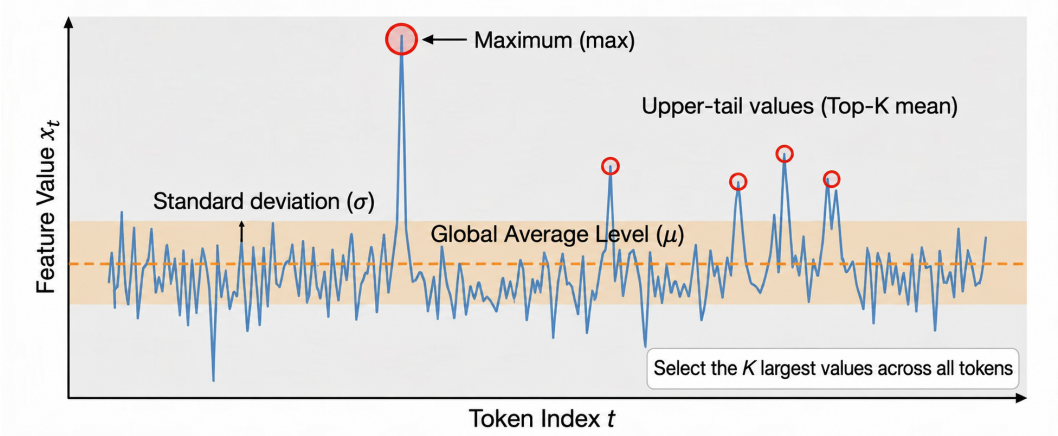

Figure 5: Schematic of the four pooling operators in Equation 5. The dashed line denotes the mean and the shaded band illustrates variation. Highlighted peaks illustrate upper-tail selection, which includes the maximum and need not occupy consecutive positions. The implemented tail size is $K = \min(T, \max(5, \lceil 0.1T \rceil))$; the curve is illustrative, not an experimental trace.

`reasoning_delta_norm`, the three KL features, and `prediction_stability`. Its pooled dimension is 45; Core has dimension 13. The family-only dimensions are 53, 61, and 33 for Conflict, Drift, and Instability. Each dimension includes response length once.

A.1 Pooling illustration

Figure 5 contrasts whole-response summaries with the largest feature values. Each operator is computed independently for each feature. Top-K averages the largest numerical values across the entire response, including its maximum; the selected positions need not be adjacent. The classifier learns how these summaries relate to hallucination labels.

A.2 Output controls

Output-gap contains CTX-minus-NOCTX differences in target logit, target log probability, target probability, entropy, and maximum probability; NOCTX-minus-CTX target log rank; JSD; symmetric KL; and the root-mean-square vocabulary-logit difference. Output-all adds each path's target negative log likelihood, target probability, entropy, maximum probability, and target log rank, yielding 19 features. The same four summary statistics and length feature yield 37 and 77 dimensions for Output-gap and Output-all.

For observed target y_t, the control distribution is read at the position immediately preceding that token. The first target is scored from the last prompt position. DualTrace's hidden-trajectory observations instead use the returned states at response-token positions, whose output logits predict the following token. Both controls retain the same observed response tokens and data splits. This convention applies to target likelihood and replay negative log likelihood (NLL), as opposed to distributions read after consuming the response token.

B Data and Training Protocol

Table 5 describes the 12 control subsets. A response is positive when at least one token has an annotated hallucination label. The supplied train/test grouping is retained; responses sharing a retrieval source cannot cross that boundary. The RAGTruth sets have 400 training and 90 test source groups per model; HalluRAG has 250 and 60. These counts describe the control subsets of the published benchmarks. External-method and evidence-path experiments use their own sets of responses available to every compared method or condition, specified with those results.

Table 5: Responses and source groups in the 12-setting control-subset evaluation. Positive test responses contain at least one annotated hallucination token.

Model	Dataset	Train	Test	Train groups	Test groups	Test +
Llama2-7B	HalluRAG	494	129	250	60	49
Llama2-7B	RAGTruth	780	181	400	90	67
Llama2-13B	HalluRAG	500	106	250	60	48
Llama2-13B	RAGTruth	783	175	400	90	58
Llama3-8B	HalluRAG	513	116	250	60	46
Llama3-8B	RAGTruth	779	167	400	90	67
Mistral-7B	HalluRAG	500	120	250	60	32
Mistral-7B	RAGTruth	795	188	400	90	77
Qwen3-8B	HalluRAG	488	109	250	60	40
Qwen3-8B	RAGTruth	798	187	400	90	53
Qwen3-14B	HalluRAG	498	121	250	60	37
Qwen3-14B	RAGTruth	791	178	400	90	68

The control-subset classifier protocol uses XGBoost 2.1.3, NumPy 1.26.4, and scikit-learn 1.3.2. The five seeds are 13, 21, 42, 87, and 101. Five-fold shuffled StratifiedGroupKFold keeps retrieval-source groups together while balancing labels within training. XGBoost uses logistic binary loss, learning rate 0.05, maximum depth 6, minimum child weight 5, row and column subsampling 0.9, and L2 regularization 1.0. The positive-class weight is the negative/positive count ratio in the corresponding training fold. Histogram tree construction runs on CPU. The external-method evaluation uses GroupKFold as specified in Appendix C.3.

Training permits up to 800 rounds with 30-round early stopping on validation AUPRC. For each seed, the median selected round count across folds determines a model trained on the full training set. Mean OOF scores determine the ensemble F1 threshold; test predictions are the mean of the five final-model scores. AUROC and AUPRC summarize rankings. Transfer instead reports the mean of the five independently evaluated seed metrics, then averages directions. The full feature matrix clips magnitudes at 10^4 and maps nonfinite entries to finite values under the frozen preprocessing policy. The controlled evidence-path extraction additionally checks that all emitted feature values are finite.

Source-group bootstrap draws source IDs with replacement and retains all their responses, preserving method pairing. Within-setting Full-versus-variant tests use Holm correction over all 11 comparisons; the external-method AUROC tests form one seven-comparison family in Table 3. The control-subset bootstrap two-sided Monte Carlo p-values use an add-one correction. The representation-analysis families and transformations are specified in Appendix E.

C Supporting Detection Results

C.1 Broad benchmark protocol and baselines

Table 1 evaluates four language models using AUROC, AUPRC, and PCC. Its RAGTruth response collections contain approximately 2,500 training and 450 test responses per model. HalluRAG uses the answerable subset with source-grouped 70%/15%/15% training/validation/test splits. The response label is positive if any token is annotated as hallucinated; severity is the fraction of hallucinated tokens. PCC is the Pearson correlation of the response score with this fraction. DualTrace uses the response-level pooling and XGBoost pipeline, with five-fold GroupKFold by source and training-OOF threshold selection. The five-seed control-subset comparisons in Table 2 and recalibrated common-response comparisons in Table 3 are separate evaluations.

Perplexity summarizes response negative log likelihood (Ren et al., 2023), while LN-Entropy summarizes length-normalized predictive entropy (Malinin & Gales, 2021). Focus weights uncertainty by token information and attention (Zhang et al., 2023). SelfCheckGPT measures consistency across sampled responses (Manakul et al., 2023); EigenScore summarizes the spectrum of their hidden-representation covariance (Chen et al., 2024). P(True) elicits a model's probability judgment of an answer's truth (Kadavath et al., 2022), and RefChecker verifies extracted claim triplets against the reference (Hu et al., 2024). AlignScore compares the response with supporting text (Zha et al., 2023);

Table 6: Signal-family ablations and score averaging, macro-averaged over the 12 settings. The best single family is selected by training-OOF AUROC. Each family detector is supervised; their equal-weight combination is fixed. Dashes denote metrics not used for the fusion comparison.

Detector	AUROC	AUPRC
Conflict only	0.9176	0.8817
Drift only	0.9111	0.8644
Instability only	0.8435	0.7659
Full without Conflict	0.9161	0.8748
Full without Drift	0.9200	0.8869
Full without Instability	0.9245	0.8952
Full DualTrace	0.9279	0.8995
Train-selected single family	0.9137	—
Equal-weight family fusion	0.9294	—

Table 7: Complete AUROC ablations on the 12 settings. C, D, and I denote Conflict, Drift, and Instability; minus denotes removal.

Model	Dataset	Full	C	D	I	$-C$	$-D$	$-I$
Llama2-7B	HalluRAG	0.9827	0.9801	0.9589	0.9125	0.9594	0.9837	0.9768
Llama2-7B	RAGTruth	0.8789	0.8716	0.8349	0.8041	0.8565	0.8801	0.8683
Llama2-13B	HalluRAG	0.8865	0.8631	0.8937	0.7899	0.9048	0.8581	0.8822
Llama2-13B	RAGTruth	0.8970	0.9036	0.8543	0.8019	0.8574	0.9022	0.8952
Llama3-8B	HalluRAG	0.9472	0.9273	0.9307	0.8376	0.9385	0.9248	0.9453
Llama3-8B	RAGTruth	0.9884	0.9830	0.9797	0.9400	0.9803	0.9852	0.9836
Mistral-7B	HalluRAG	0.9627	0.9446	0.9592	0.8956	0.9698	0.9510	0.9634
Mistral-7B	RAGTruth	0.8676	0.8491	0.8530	0.7565	0.8578	0.8467	0.8662
Qwen3-8B	HalluRAG	0.9029	0.8884	0.8822	0.8196	0.8812	0.9040	0.8971
Qwen3-8B	RAGTruth	0.9382	0.9254	0.9238	0.8591	0.9223	0.9265	0.9362
Qwen3-14B	HalluRAG	0.9450	0.9456	0.9231	0.8172	0.9208	0.9463	0.9421
Qwen3-14B	RAGTruth	0.9384	0.9290	0.9398	0.8881	0.9443	0.9311	0.9373

the benchmark implementation uses AlignScore-large with mean pooling across context chunks. ReDeEP provides context and parametric contribution scores (Sun et al., 2025). Table 1 evaluates it under the broad benchmark protocol; Table 3 evaluates the explicit train-calibrated scoring rule and XGBoost adaptation on common response sets.

C.2 Feature ablations and score averaging

Table 6 gives the AUROC and AUPRC summary accompanying Figure 2. Tables 7 and 8 report the complete feature-set comparisons, including all removal variants and fixed reduced representations. Table 9 reports the score-averaging test in every setting. The best single family is chosen from its mean training-OOF AUROC before comparing test predictions. Fusion averages the three separately trained family scores without further calibration of combination weights.

Figure 6 supplements the family-level comparison with finer-grained removals. Llama2-13B has its largest plotted removal loss for the LogitLens group, while Llama3-8B is more sensitive to LSD and LogitLens than to ICR. These within-model contrasts describe different dependencies on the observed feature groups. The configurations follow the broad benchmark in Table 1; Figure 2 summarizes the separate 12-setting control-subset evaluation.

C.3 External-method implementations

Table 3 uses a unified RAGTruth evaluation collection of 2,965 responses from 128 source groups per backbone. All comparisons within a backbone share responses, labels, and a single set of DualTrace predictions. These collections are separate from the control subsets in Table 5. The shared XGBoost protocol uses five-seed ensembles and five-fold GroupKFold, with source groups kept disjoint between training and evaluation. DualTrace's AUROC/AUPRC is 0.8924/0.8540 on

Table 8: Complete AUROC ablations on the 12 settings. Compact, Core, and Projection are fixed subsets defined in Appendix A. Gap and All-out denote Output-gap and Output-all.

Model	Dataset	Full	Compact	Core	Projection	Gap	All-out
Llama2-7B	HalluRAG	0.9827	0.9696	0.9640	0.9434	0.9334	0.9360
Llama2-7B	RAGTruth	0.8789	0.8515	0.7121	0.8096	0.7615	0.7523
Llama2-13B	HalluRAG	0.8865	0.8890	0.8114	0.8495	0.8412	0.8473
Llama2-13B	RAGTruth	0.8970	0.8774	0.7881	0.8466	0.8559	0.8736
Llama3-8B	HalluRAG	0.9472	0.9354	0.8071	0.9280	0.8769	0.8851
Llama3-8B	RAGTruth	0.9884	0.9854	0.9416	0.9631	0.9246	0.9355
Mistral-7B	HalluRAG	0.9627	0.9602	0.8984	0.9414	0.9577	0.9599
Mistral-7B	RAGTruth	0.8676	0.8424	0.7810	0.8188	0.8343	0.8208
Qwen3-8B	HalluRAG	0.9029	0.8996	0.8065	0.8649	0.8736	0.8679
Qwen3-8B	RAGTruth	0.9382	0.9050	0.8417	0.9150	0.8519	0.8775
Qwen3-14B	HalluRAG	0.9450	0.9517	0.8633	0.9012	0.8803	0.9048
Qwen3-14B	RAGTruth	0.9384	0.9299	0.8672	0.9148	0.9283	0.9297

Table 9: Complete score-averaging comparison. Intervals are for equal-weight fusion minus the training-selected single family.

Setting	Selected family	Single	Fusion	Δ 95% CI
Llama2-7B / HalluRAG	Conflict	0.9801	0.9827	$[-0.013, 0.018]$
Llama2-7B / RAGTruth	Conflict	0.8716	0.8849	$[-0.021, 0.048]$
Llama2-13B / HalluRAG	Conflict	0.8631	0.9055	$[0.002, 0.084]$
Llama2-13B / RAGTruth	Drift	0.8543	0.8899	$[0.011, 0.062]$
Llama3-8B / HalluRAG	Drift	0.9307	0.9450	$[-0.013, 0.047]$
Llama3-8B / RAGTruth	Conflict	0.9830	0.9890	$[-0.003, 0.018]$
Mistral-7B / HalluRAG	Conflict	0.9446	0.9645	$[-0.010, 0.051]$
Mistral-7B / RAGTruth	Conflict	0.8491	0.8657	$[-0.012, 0.047]$
Qwen3-8B / HalluRAG	Conflict	0.8884	0.8960	$[-0.034, 0.045]$
Qwen3-8B / RAGTruth	Conflict	0.9254	0.9310	$[-0.012, 0.023]$
Qwen3-14B / HalluRAG	Conflict	0.9456	0.9437	$[-0.023, 0.020]$
Qwen3-14B / RAGTruth	Conflict	0.9290	0.9553	$[0.006, 0.050]$

Llama3-8B and 0.8085/0.7620 on Llama2-7B. Paired intervals use 10,000 source-group bootstrap repetitions with seed 42; Holm correction covers all seven AUROC comparisons. The reported 95% intervals are marginal, whereas the adjusted tests control the seven-comparison family.

Each method's original scoring rule or classifier is labeled native; feature-based XGBoost adaptations are labeled separately. ReDeEP uses the official feature extractor at commit `4d081915`. Its paper-scoring evaluation fixes the top external and parametric components to two and five and uses $\alpha = 1.2$, with layer selection and min–max calibration restricted to training data. The feature adaptation uses the same DualTrace pooling and five-seed XGBoost protocol.

LUMINA uses the equations at official commit `c43ff41d`, with memory-bounded computation, $\lambda = 0.5$, and top-100 support for maximum mean discrepancy (MMD), its distributional distance. Its native response score is the arithmetic mean of token scores. We also evaluate the extracted components with the shared XGBoost classifier. The native score requires no detector-label training.

CORTEX is a paper reimplementation, using token-level labels to fit a 256–128–1 multilayer perceptron (MLP) with ReLU and dropout 0.1, followed by the paper's label-persistence smoothing. The reported native result is a five-seed ensemble. RAGLens uses official commit `e10ed582`, the public layer-15 SAE, response max pooling, train-only mutual-information selection of 1,000 features, and a native generalized additive model. Its public SAE training checkpoint and the activation model are compatible but not identical. The RAGLens feature adaptation replaces its GAM with the shared five-seed XGBoost classifier. Table 3 reports results computed with these recipes on the common response sets.

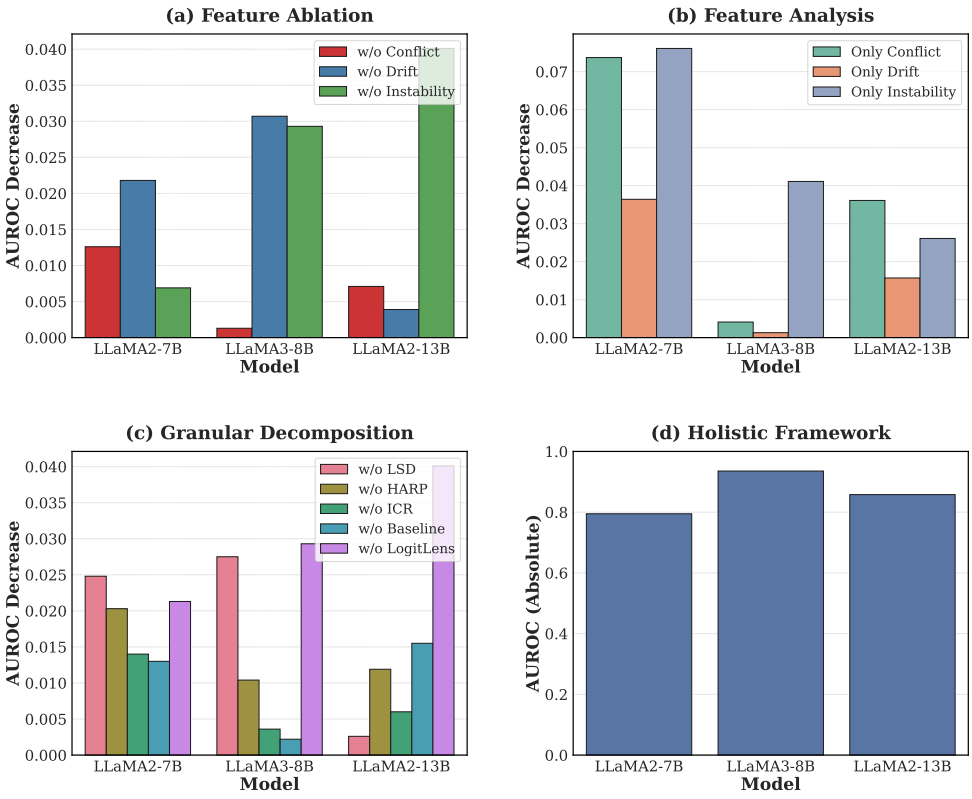

Figure 6: Component-level analysis under the broad RAGTruth benchmark protocol. Panels (a)–(c) show Full-minus-variant AUROC; panel (d) shows Full AUROC. In (c), labels name internal feature groups: LSD denotes displacement statistics, HARP the output-head projection features, ICR the alignment/JSD group, Baseline the base context/path features, and LogitLens the layerwise predictive features.

Table 10: Evidence-path controls, equal-weighted over six settings. All four conditions use common responses and unchanged labels. Each row retrains classifiers with the same protocol on the corresponding path features.

Compared paths	Full	Output-all	Difference	Full higher
CTX / NOCTX	0.8940	0.8667	0.0273	5/6
Random / NOCTX	0.8597	0.8468	0.0128	2/6
CTX / random	0.8980	0.8599	0.0381	5/6
CTX+random / NOCTX	0.8969	0.8647	0.0322	6/6

C.4 Evidence-path controls

The evidence-path experiment changes the documents supplied to the two replay paths while preserving the response and its label. Its conditions are relevant-context/NOCTX, random-context/NOCTX, relevant-context/random-context, and relevant-plus-random-context/NOCTX. Random-context documents and document boundaries are set before training. Within each model–dataset setting, all four conditions use the responses available in every condition, and both classifiers use the same training/test grouping. A fixed 4,096-token extraction limit and responses exceeding hardware capacity determine availability before this intersection. The comparison therefore evaluates the same responses across all four conditions. Tables 10 and 11 provide the macro and complete setting results.

Table 11: Complete evidence-path AUROC matrix. H: HalluRAG; R: RAGTruth. Intervals are Full minus Output-all on identical shared-response sets.

Model	Data	Paths	Full	Output-all	Δ 95% CI
Llama3-8B	H	CTX / NOCTX	0.9587	0.9048	$[0.000, 0.126]$
Llama3-8B	H	Random / NOCTX	0.9794	0.8873	$[0.020, 0.188]$
Llama3-8B	H	CTX / random	0.9579	0.8889	$[0.012, 0.141]$
Llama3-8B	H	CTX+random / NOCTX	0.9698	0.8540	$[0.029, 0.228]$
Llama3-8B	R	CTX / NOCTX	0.9178	0.9118	$[-0.006, 0.018]$
Llama3-8B	R	Random / NOCTX	0.8588	0.8707	$[-0.036, 0.012]$
Llama3-8B	R	CTX / random	0.9119	0.9113	$[-0.012, 0.013]$
Llama3-8B	R	CTX+random / NOCTX	0.9166	0.9114	$[-0.008, 0.019]$
Mistral-7B	H	CTX / NOCTX	0.9315	0.9140	$[-0.021, 0.057]$
Mistral-7B	H	Random / NOCTX	0.9384	0.9016	$[-0.004, 0.080]$
Mistral-7B	H	CTX / random	0.9412	0.8998	$[0.001, 0.085]$
Mistral-7B	H	CTX+random / NOCTX	0.9317	0.9055	$[-0.013, 0.066]$
Mistral-7B	R	CTX / NOCTX	0.8435	0.8242	$[0.000, 0.039]$
Mistral-7B	R	Random / NOCTX	0.8118	0.8288	$[-0.041, 0.007]$
Mistral-7B	R	CTX / random	0.8478	0.8388	$[-0.009, 0.028]$
Mistral-7B	R	CTX+random / NOCTX	0.8444	0.8327	$[-0.009, 0.033]$
Qwen3-8B	H	CTX / NOCTX	0.8636	0.7955	$[-0.178, 0.494]$
Qwen3-8B	H	Random / NOCTX	0.7175	0.7321	$[-0.539, 0.581]$
Qwen3-8B	H	CTX / random	0.8831	0.7727	$[-0.094, 0.468]$
Qwen3-8B	H	CTX+random / NOCTX	0.8604	0.8393	$[-0.158, 0.256]$
Qwen3-8B	R	CTX / NOCTX	0.8488	0.8498	$[-0.017, 0.015]$
Qwen3-8B	R	Random / NOCTX	0.8520	0.8604	$[-0.025, 0.008]$
Qwen3-8B	R	CTX / random	0.8459	0.8478	$[-0.019, 0.016]$
Qwen3-8B	R	CTX+random / NOCTX	0.8584	0.8452	$[-0.002, 0.028]$

C.5 Replay likelihood and faithful agreement

Replaying a contextual response without documents can make its tokens unlikely. The replay-control experiment tests how much detection can be explained by this effect. It uses a separately fixed protocol with seed 42, the same feature families, four-statistic pooling, and training-OOF threshold selection. Artifact features include NOCTX target likelihood, target rank, entropy, maximum probability, forced-prefix mean negative log likelihood, and indicators of low-probability forced tokens. Artifact-only predicts hallucination from these likelihood measures; matching compares responses with similar replay characteristics. This single-run evaluation is separate from the five-seed main evaluation.

Across the 12 settings, Full exceeds Artifact-only by a mean AUROC of 0.0580 and reaches a mean AUROC of 0.8896 after artifact matching. Replay likelihood is predictive, but Full remains predictive among responses matched on these likelihood measures. Table 12 reports every setting.

The agreement control examines faithful responses when the model can answer correctly without retrieval. A faithful agreement case has a naturally generated NOCTX answer that agrees with the supported answer, according to manually checked labels. The pooled false-positive rate (FPR) is 17.55% among 906 faithful agreement cases (Wilson 95% CI [15.21, 20.16]), compared with 15.72% among 229 faithful non-agreement cases (CI [11.58, 21.00]). FPR is the fraction flagged by the fixed detector within each faithful subset. In contrast, Figure 8 reports the fraction of hallucinated responses in each E/I group, without applying the detector. Table 12 gives setting-specific agreement counts and FPRs.

C.6 Transfer

Table 13 reports all 18 transfer directions. Source training and its OOF predictions determine model fitting and thresholds. Same-dataset cross-model comparisons retain source-group separation from target test responses. The output control and both DualTrace variants use the same transfer directions and feature definitions.

Table 12: Replay and agreement controls under their separately frozen single-run protocol. H: HalluRAG; R: RAGTruth. Matched denotes Full AUROC after artifact matching. Agree. n and FPR refer only to faithful agreement cases; FPR is a percentage.

Model	Data	Test n	Full	Artifact	Matched	Agree. n	FPR
Llama2-13B	H	106	0.8822	0.8297	0.8568	49	38.8
Llama2-13B	R	175	0.8939	0.8506	0.8317	93	23.7
Llama2-7B	H	129	0.9875	0.9217	0.9617	62	1.6
Llama2-7B	R	181	0.8727	0.7701	0.8670	85	47.1
Llama3-8B	H	116	0.9404	0.8441	0.9556	53	22.6
Llama3-8B	R	167	0.9879	0.9194	0.9771	80	1.2
Mistral-7B	H	120	0.9670	0.9545	0.9570	71	23.9
Mistral-7B	R	188	0.8529	0.7839	0.7350	84	14.3
Qwen3-14B	H	121	0.9519	0.8912	0.9503	73	4.1
Qwen3-14B	R	178	0.9345	0.9278	0.8737	87	6.9
Qwen3-8B	H	109	0.9007	0.8388	0.8088	55	18.2
Qwen3-8B	R	187	0.9373	0.8810	0.9010	114	14.0

Table 13: Complete transfer matrix. Entries are mean AUROC over five seeds. Source-only OOF predictions determine calibration; target test labels are used for evaluation.

Source	Target	Full	Compact	Output-all
Llama2-7B / RAGTruth	Llama2-7B / HalluRAG	0.9763	0.9602	0.9437
Llama2-7B / HalluRAG	Llama2-7B / RAGTruth	0.8785	0.8465	0.7573
Llama2-13B / RAGTruth	Llama2-13B / HalluRAG	0.9035	0.9054	0.8107
Llama2-13B / HalluRAG	Llama2-13B / RAGTruth	0.8906	0.8566	0.8386
Llama3-8B / RAGTruth	Llama3-8B / HalluRAG	0.9073	0.9206	0.8717
Llama3-8B / HalluRAG	Llama3-8B / RAGTruth	0.9816	0.9652	0.9223
Mistral-7B / RAGTruth	Mistral-7B / HalluRAG	0.9513	0.9290	0.9536
Mistral-7B / HalluRAG	Mistral-7B / RAGTruth	0.8554	0.8479	0.8275
Qwen3-8B / RAGTruth	Qwen3-8B / HalluRAG	0.9155	0.9048	0.8402
Qwen3-8B / HalluRAG	Qwen3-8B / RAGTruth	0.9154	0.8859	0.8875
Qwen3-14B / RAGTruth	Qwen3-14B / HalluRAG	0.9450	0.9488	0.8732
Qwen3-14B / HalluRAG	Qwen3-14B / RAGTruth	0.9436	0.9211	0.9293
Llama3-8B / RAGTruth	Mistral-7B / RAGTruth	0.8517	0.8585	0.8375
Llama3-8B / RAGTruth	Qwen3-8B / RAGTruth	0.9145	0.8918	0.8985
Qwen3-8B / RAGTruth	Qwen3-14B / RAGTruth	0.9351	0.9288	0.9200
Llama3-8B / HalluRAG	Mistral-7B / HalluRAG	0.9493	0.9382	0.9583
Llama3-8B / HalluRAG	Qwen3-8B / HalluRAG	0.9246	0.8851	0.8729
Qwen3-8B / HalluRAG	Qwen3-14B / HalluRAG	0.9469	0.9465	0.9039

C.7 TIMING SCOPE

Figure 7 separates the cost of the two forward passes from the cost of the complete detection pipeline on fixed responses.

The forward-only benchmark uses 100 seed-selected, class-balanced test responses per model, ten warmup responses, and five timing repetitions per response. CTX-output-only runs the contextual forward without hidden-state collection; the dual path runs CTX and NOCTX with hidden-state collection. Reported ratios are medians of per-response latency ratios. Tokenization, feature pooling, classification, model loading, and disk I/O are excluded from this benchmark.

The separate full-detector benchmark includes raw-text tokenization, two teacher-forced forwards, feature equations, response pooling, and a 100-round XGBoost classifier. It excludes model loading, projection initialization, detector training, and disk I/O. The median/95th-percentile times are 0.230/0.567 seconds for Llama3-8B, 0.233/0.443 for Mistral-7B, and 0.281/0.582 for Qwen3-8B. Peak allocated memory maxima are 18.25, 16.60, and 18.54 GB, respectively (decimal units). These are fixed-response detection costs; generation and retrieval are outside the measured interval.

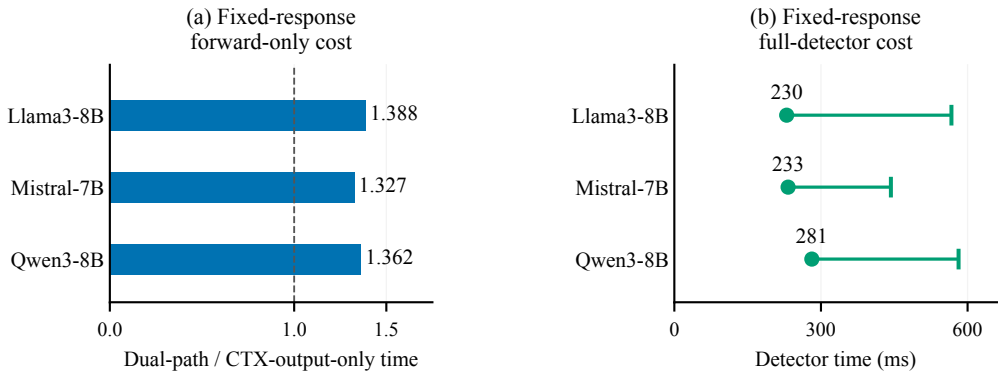

Figure 7: Fixed-response detection cost on 100 held-out responses per model. (a) Median per-response dual-path/CTX-output-only forward-time ratios. (b) Full-detector median times (circles) and 95th percentiles (vertical end marks), including the specified 100-round classifier. Spans show runtime variation, not confidence intervals.

On the corresponding 100 Llama3-8B responses, CORTEX representation extraction takes 0.171 seconds at the median and LUMINA tokenization/forwards/equations/pooling takes 0.295 seconds. The CORTEX measurement ends at representation extraction, whereas DualTrace's 0.230-second measurement includes its 100-round classifier. Each time therefore characterizes the specified portion of its detection pipeline.

D Activation-Replacement Protocol

We choose transformer blocks on a discovery split and evaluate activation replacement on separate held-out responses. Selected zero-based blocks are 23, 27, and 31 for Llama3-8B and Mistral-7B, and 17, 22, and 35 for Qwen3-8B. Confirmation uses 60 faithful and 60 hallucinated RAGTruth responses per model. The joint representation–intervention analysis contains 359 of these 360 responses because one Mistral response lies outside the response set shared by the four context conditions.

At a selected block, we replace the NOCTX residual states at response positions with $(1-\alpha)h^{\mathrm{NOCTX}}+\alpha h^{\mathrm{CTX}}$, using $\alpha \in \{0, 0.25, 0.5, 0.75, 1\}$. Each donor path supplies hidden states computed on the same fixed response. Controls use a random-context donor into NOCTX and reverse NOCTX-to-CTX replacement. For donor b at block ℓ, the measured effect is

$$A_{b,\ell}(r) = \frac{1}{T-1}\sum_{t=2}^{T}\left[\log p_{b,\ell,\alpha=1}(y_t \mid y_{<t}) - \log p_{\mathrm{base}}(y_t \mid y_{<t})\right]. \tag{9}$$

The base path is NOCTX for forward replacements and CTX for the reverse control. The effect $A_{b,\ell}$ is a signed log-probability difference. The score used for the representation–intervention comparison is $J(r) = |\mathcal{L}|^{-1}\sum_{\ell\in\mathcal{L}}[A_{\mathrm{CTX},\ell}(r) - A_{\mathrm{random},\ell}(r)]$. This endpoint measures the difference between context-donor and random-donor probability effects under the specified replacement operation. At the final block, full replacement supplies donor states directly to the final normalization and output head, providing a readout reference. The non-final-block analysis measures effects that pass through subsequent transformer blocks.

Table 14 summarizes the confirmation controls. At $\alpha = 0$, the mean absolute change is zero in all three models. The full interpolation grid and both classes enter the layer-averaged comparisons. Figure 4 examines each selected block separately, including both non-final and final blocks; this layer-by-layer analysis is exploratory.

Table 14: Held-out activation replacement: CTX-donor effect relative to random-context and reverse controls, averaged over the layers chosen on separate discovery data. Effects measure changes in fixed-response log probability.

Model	n	Layers	Main−random	95% CI	Main−reverse
Llama3-8B	120	23,27,31	1.4655	$[1.364, 1.563]$	2.9292
Mistral-7B	120	23,27,31	1.1376	$[1.059, 1.216]$	2.3042
Qwen3-8B	120	17,22,35	0.6168	$[0.564, 0.671]$	2.0083

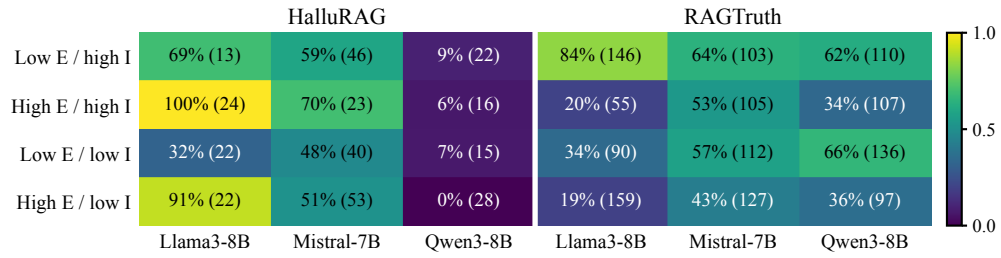

Figure 8: Context sensitivity E and prediction disagreement I across six settings ($n = 1,671$). Cells give hallucination rates and counts in parentheses. Equation 7 and fixed 0.5 cutoffs define the groups; colors show rates from 0 to 1.

E Context Sensitivity and Prediction Agreement: Analysis Details

Figure 2a shows an illustrative Llama3-8B/RAGTruth setting with 255 faithful and 195 hallucinated responses. Class-conditional Gaussian kernel density estimates use Scott's bandwidth and eight contour levels at 16–90% of each class's peak density, evaluated on a 91×91 grid over $[0, 1]^2$. These contours display density, not confidence regions or detector boundaries. Figure 8 supplies the group rates for all six settings.

The analysis uses the six CTX–NOCTX settings with shared responses from the evidence-path control matrix. Training splits calibrate empirical percentiles and robust scales; test splits contain 1,671 responses (324 HalluRAG and 1,347 RAGTruth). HalluRAG validation responses are unused. The responses shared with the intervention experiment include 120 Llama3-8B, 119 Mistral-7B, and 120 Qwen3-8B responses; one Mistral response lies outside the response set shared across context conditions. These sample sets are specific to this analysis and differ from the full 12-setting detection evaluation.

For training measurements $a_1, \ldots, a_n$, the smoothed midrank percentile assigned to x is $(\#\{a_i < x\} + \frac{1}{2}\#\{a_i = x\} + \frac{1}{2})/(n + 1)$. Context sensitivity E averages the output-JSD and trajectory-divergence percentiles; prediction disagreement I is the late-to-final KL percentile. The fixed 0.5 cutoff for E is a descriptive boundary; for I, it corresponds to the training median. Test labels determine neither score orientation nor cutoffs. The hallucination rate in each group is its fraction of labeled hallucinated responses, independently of any detector threshold.

Layer-stage measurements use response means, with $\log(1 + x)$ applied to KL, displacement magnitude (velocity), and direction change (acceleration) before training-median/IQR scaling; entropy and prediction stability retain their original scales before standardization. If the training IQR is zero, the implementation uses training standard deviation, with unit scale as a final fallback. The combined late-layer change averages standardized velocity and acceleration and subtracts standardized prediction stability, followed by training-based standardization of the composite. All intervals resample source groups 10,000 times. Macros equally weight the evaluated settings or models. Holm correction covers four quadrant contrasts, eleven layer-stage measurements, and three primary aggregate intervention correlations, respectively. This analysis plan was fixed before computing these additional summaries of existing experimental results; the subsequent layer-by-layer analysis is exploratory.

Table 15: All four group contrasts across six settings. E is context sensitivity and I is prediction disagreement. The reference is high E/low I; values are hallucination-rate differences in percentage points. The last row combines the low/low and high/high groups.

Group	Difference	95% CI	p_{Holm}
Low E, low I	$+0.63$	$[-6.08, 7.66]$	0.8707
Low E, high I	$+17.75$	$[10.35, 25.20]$	0.0008
High E, high I	$+7.14$	$[1.10, 13.17]$	0.0684
Low/low and high/high combined	$+4.26$	$[-1.26, 9.89]$	0.2488

Table 16: Complete set of layer-stage measurements. Values are equal-weight macro differences between hallucinated and faithful responses, in training-IQR units after the stated transformations.

Measurement	Difference	95% CI	p_{Holm}
Early-to-final KL	-0.309	$[-0.510, -0.157]$	0.0022
Middle-to-final KL	-0.153	$[-0.259, -0.055]$	0.0112
Late-to-final KL	$+0.280$	$[0.151, 0.397]$	0.0022
Early entropy	-0.308	$[-0.482, -0.088]$	0.0300
Middle entropy	-0.037	$[-0.136, 0.066]$	1.0000
Final entropy	$+0.400$	$[0.266, 0.545]$	0.0022
Late velocity	-0.012	$[-0.122, 0.098]$	1.0000
Late acceleration	$+0.157$	$[0.006, 0.305]$	0.2030
Prediction stability	$+0.007$	$[-0.140, 0.132]$	1.0000
Late-minus-early log-KL	$+0.316$	$[0.193, 0.437]$	0.0022
Combined late-layer change	$+0.044$	$[-0.083, 0.203]$	1.0000

The three primary aggregate intervention correlations are context sensitivity E ($\rho = 0.625$, 95% CI [0.549, 0.691], $p_{\text{Holm}} = 0.0006$), prediction disagreement I ($\rho = -0.308$, 95% CI $[-0.399, -0.209]$, $p_{\text{Holm}} = 0.0006$), and their product ($\rho = 0.146$, 95% CI [0.037, 0.252], $p_{\text{Holm}} = 0.0096$). Model-specific context-sensitivity correlations are 0.798 [0.710, 0.861], 0.890 [0.838, 0.922], and 0.187 $[-0.013, 0.376]$ for Llama3-8B, Mistral-7B, and Qwen3-8B. The exploratory averages over the selected non-final blocks yield 0.797 [0.706, 0.860], 0.890 [0.838, 0.923], and -0.370 $[-0.526, -0.198]$, respectively. These estimates characterize changes in fixed-response log probability at the evaluated layers.

F LIMITATIONS

DualTrace requires access to model hidden states and output distributions, so it does not directly apply to text-only APIs. Training requires labeled responses, and detection adds a NOCTX forward pass and feature computation. Transfer is evaluated on RAGTruth and HalluRAG with the selected backbones; performance on other domains and architectures remains to be established. The representation analyses describe selected measurements, while activation replacement tests changes in fixed-response probability rather than whether generated answers become more faithful.